This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3134455



# Adversarial Shape Learning for Building Extraction in VHR Remote Sensing Images

Lei Ding, Hao Tang, Yahui Liu, Yilei Shi, *Member, IEEE,* Xiao Xiang Zhu, *Fellow, IEEE* and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Building extraction in VHR RSIs remains a challenging task due to occlusion and boundary ambiguity problems. Although conventional convolutional neural networks (CNNs) based methods are capable of exploiting local texture and context information, they fail to capture the shape patterns of buildings, which is a necessary constraint in the human recognition. To address this issue, we propose an adversarial shape learning network (ASLNet) to model the building shape patterns that improve the accuracy of building segmentation. In the proposed ASLNet, we introduce the adversarial learning strategy to explicitly model the shape constraints, as well as a CNN shape regularizer to strengthen the embedding of shape features. To assess the geometric accuracy of building segmentation results, we introduced several object-based quality assessment metrics. Experiments on two open benchmark datasets show that the proposed ASLNet improves both the pixel-based accuracy and the object-based quality measurements by a large margin. The code is available at: *https://github.com/ggsDing/ASLNet*.**

***Index Terms*—Building Extraction, Generative Adversarial Networks (GANs), Image Segmentation, Convolutional Neural Network, Deep Learning, Remote Sensing**

## I. Introduction

Shape is an important pattern in the process of visual recognition. Direct modeling of shape patterns in images is challenging since it requires a high-level abstract of the object contours. Among the real-world applications of image recognition techniques, building extraction in very high resolution (VHR) remote sensing images (RSIs) is one of the most interesting and challenging tasks that can benefit greatly from learning the shape patterns. It is important for a wide variety of applications, such as land-cover mapping, urban resources management, detection of illegal constructions, etc.

Conventional building extraction algorithms are based on handcrafted features that often fail to model high-level context information and are highly dependent on parameters. Recently, with the emergence of convolutional neural networks (CNNs) and their applications in semantic segmentation tasks (e.g., vehicle navigation [1], scene parsing [2], medical image segmentation [3]), a large research interest has been focused on adapting these CNN models to building extraction in VHR RSIs. The CNN-based building extraction methods employ stacked convolution operations to extract the intrinsic content information of images, thus they are more effective in exploiting the context information while they are less sensitive to domain changes. A variety of CNN designs for the semantic segmentation of buildings have been introduced with good results [4], [5].

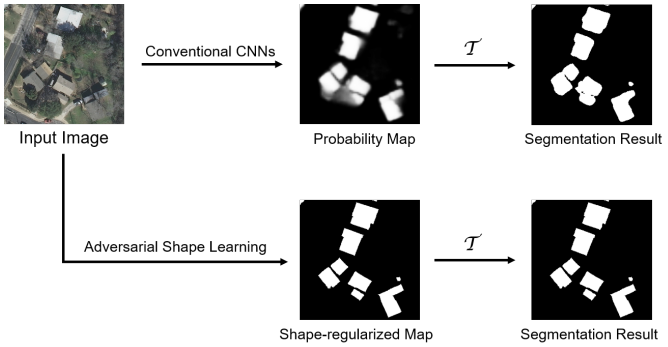



Fig. 1: Illustration of the benefits of the proposed shape learning. Conventional CNN models lead to boundary ambiguity problems, whereas the proposed method produces shape-regularized results.

However, some critical challenges in building extraction remain unsolved even with the use of the recent CNN-based methods. First, occlusions (caused by trees and shadows) and intra-class diversity are common problems in VHR RSIs, which often cause fragmentation and incomplete segmentation. Second, it is common to have boundary ambiguity problems. Due to the effects of shadows and building profiles, an accurate localization of the building boundaries is difficult (especially in the low-contrast areas). Conventional CNN-based methods produce ambiguous probability values in these areas, which often cause rounded or uneven building boundaries after thresholding. Last but not least, the segmentation maps generally suffer from over-segmentation and under-segmentation errors caused by these fragmentation and boundary-adhesion problems. Due to these limitations, post-processing algorithms are often required to optimize the building extraction results [6], [7].

L. Ding, Y. Liu, and L. Bruzzone are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (E-mail: dinglei14@outlook.com, hao.tang@unitn.it, yahui.liu@unitn.it, lorenzo.bruzzone@unitn.it).

H. Tang is with the Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland. (E-mail: hao.tang@vision.ee.ethz.ch).

Y. Shi is with the Chair of Remote Sensing Technology, Technical University of Munich (TUM), 80333 Munich, Germany (E-mail: yilei.shi@tum.de).

X. Zhu is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany, and also with the Data Science in Earth Observation (SiPEO, formerly Signal Processing in Earth Observation), Technical University of Munich (TUM), 80333 Munich, Germany (E-mail: xiaoxiang.zhu@dlr.de).

This work is supported by the scholarship from China Scholarship Council under the grant NO.201703170123.

Another important issue is that previous works on CNN-based building extraction pay more attention to the extraction of texture and context information in RSIs, whereas the explicit modeling of building shapes has rarely been studied. In most cases, buildings in VHR RSIs are compact and rectangular objects with sharp edges and corners. Their rectangularity is very discriminative compared to other ground objects. Learning this shape prior is beneficial for not only inpainting the occluded building parts but also reducing the boundary ambiguities and regularizing the segmentation results. An example is shown in Fig. 1 to illustrate the limitations of conventional CNNs and the benefits of the shape modelling.

In this work, we aim to address the previously mentioned issues and to improve the extraction of buildings by introducing an adversarial learning of their shape information. In greater detail, the main contributions of this work are as follows:

1) Proposing an adversarial shape learning network (ASLNet) to learn shape-regularized building extraction results. It includes a shape discriminator to exclude redundant information and focus on modelling the shape information, as well as a shape regularizer to enlarge the receptive fields (RFs) and explicitly model the local shape patterns.
2) Designing three object-based quality assessment metrics to quantitatively evaluate the geometric properties of the building extraction results. These metrics take into account both the under-segmentation and over-segmentation problems and the shape errors of the predicted building items.
3) Achieving the state-of-the-art performance on the Inria and Massachusetts building extraction benchmark datasets. Without using sophisticated backbone CNN architectures or post-processing operations, the proposed ASLNet outperforms all the compared literature methods in both pixel-based and object-based metrics.

The remainder of this paper is organized as follows. Section II introduces the related works on building extraction and adversarial learning. Section III illustrates the proposed ASLNet. Section IV describes the implementation details and the experimental settings. Section V presents the results and analyzes the effect of the proposed method. Section VI draws the conclusions of this study.

## II. Related Work

### A. CNN-based Building Extraction

Literature work focus on CNN for building extraction can be roughly divided into three types based on the studied perspectives: supervisions, architecture designs and the development of post-processing algorithms. To begin with, while binary ground truth maps are widely used to compute the segmentation losses, several papers have explored the use of other kinds of supervisions. In [8], the supervision of signed distance map (SDM) is introduced to highlight the difference between building boundaries and inner structures. In [9] signed distance labels are also introduced but in the form of classification supervision. This SDM has also been used in [10] as an auxiliary supervision.

Most CNN models for building extraction are variants of the well-known architectures for image classification and semantic segmentation. In [4], the ResUNet has been introduced for building extraction from VHR RSIs, which combines ResNet [11] with the UNet [3] structure. The MFCNN in [6] is also a symmetric CNN with ResNet as the feature extractor, whereas it contains more sophisticated designs (such as dilated convolution units and pyramid feature fusion). In [12], a Siamese UNet with two branches is designed to extract buildings from different spatial scales. In [13] a hybrid network with multiple sub-nets is introduced to exploit information from the multi-source input data. In [5], the MAPNet is proposed, which is a HRNet-like architecture with multiple feature encoding branches and channel attention designs. In [14], the global multi-scale encoder-decoder network (GMEDN) is proposed, which consists of a UNet-like network and a non-local modelling unit.

Since conventional CNN models only produce coarse segmentation results, post-processing operations are often required to obtain detailed results. In [4], guided filters are used to optimize the segmented building boundaries and to remove noise. In [7] and [15], regularization algorithms are developed to refine the segmentation maps. These algorithms perform object-based analysis on the edges and junction points to generate building-like polygons. In [6], a regularization algorithm is designed based on morphological operations on the rotated segmentation items. In [16], a graph-based conditional random field (CRF) model is combined with the segmentation network to refine the building boundaries.

### B. Adversarial Learning

*1) Generative Adversarial Networks (GANs) [17]:* GANs typically consist of two important components: a generator and a discriminator. The aim of the generator is to generate realistic results from the input data, while the discriminator is used to distinguish between the real data and the generated one. Since the discriminator is also a CNN, it is capable of learning the intrinsic differences between the real and fake data, which can hardly be modeled by human-defined algorithms. Therefore, the GANs have been widely used for a variety of complex tasks in the computer vision field, such as image generation [18], [19], [20], [21], semantic segmentation [22], [23], object detection [24], [25], depth estimation [26], and image/action recognition [27], [28].

*2) Adversarial Learning for Building Extraction:* Several literature works have introduced the adversarial learning strategy for building extraction. The segmentation model can be seen as a generative network, thus the building segmentation results can be learned in an adversarial manner by employing a CNN discriminator. The work in [29] is an early attempt on using the adversarial learning for building extraction. It forwards the masked input RSIs to the discriminator and uses an auto-encoder to reconstruct it. In [30] the GAN has been used to generate synthetic depth maps, thus improving the accuracy of building segmentation. In [31] the generative adversarial learning is introduced to improve the accuracy of building segmentation by employing a discriminator to distinguish whether the segmentation map is the ground truth (GT)

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3134455



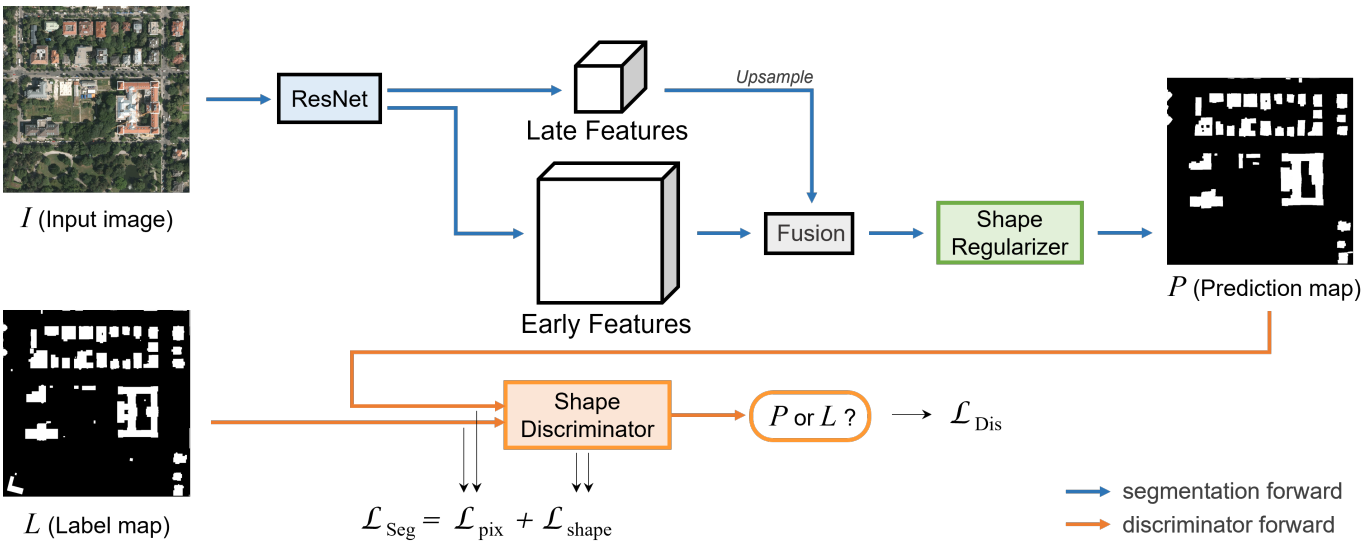



Fig. 2: Architecture of the proposed Adversarial Shape Learning Network (ASLNet) for building extraction. We designed an explicit shape regularizer (SR) to model the shape features, and a shape discriminator (SD) to guide the segmentation network. The SD discriminates whether its input is the prediction map ($P$) or the label map ($L$).

map or the segmentation results. In [32], a multi-scale L1 loss is calculated from the discriminator to train the segmentation network. In [33], a conditional Wasserstein GAN with gradient penalty (cwGAN-GP) is proposed for building segmentation, which combines the conditional GAN and Wasserstein GAN.

In general, the literature papers on the use of adversarial learning for building extraction combine the segmentation maps and the RSIs as input data to the discriminator, whereas they do not exploit the shape of segmented items.

### C. CNN-based Shape Modelling

There is a limited number of papers on CNN-based modelling of 2D shapes. To begin with, the work in [34] shows that CNNs can recognize shapes in binary images with high accuracy. In [35], the modelling of shape information is studied for the segmentation of kidneys from ultrasound scan images. In this work, a CNN auto-encoder is introduced to regularize the CNN output, which is pre-trained to recover the intact shape from randomly corrupted shapes. The shape regularization network is trained by three loss terms that measure the distance between the input segmentation map, regularized segmentation map, and the ideal segmentation map. In [36], a gated shape CNN is proposed for the semantic segmentation. It contains an explicit shape stream that deals with the object boundary information.

Several works use binary mask features to preserve and model the shape information. In [37], the shape priors are modeled to improve the instance segmentation. The label masks are cluttered to generate class-wise shape priors. These priors are then weighted by a learnt vector of parameters to estimate the coarse instance region. In [38], a shape-variant convolution is proposed for the semantic segmentation. It uses a novel paired convolution to learn context-dependent masks to limit the receptive fields (RFs) on interested image regions. In [39], the modeling of object contour polygons is studied for the instance segmentation. The polygons are first generated with a segmentation CNN and then transformed in a transformer network to fit to the object contours.

To the best of our knowledge, there is no existing work that explicitly models shape constraints for the segmentation of remote sensing images.

## III. Adversarial Shape Learning Network

Typical CNN models [4], [6] for building segmentation exploit only the local texture and context information, thus the fragmentation and boundary ambiguity problems remain unsolved. Since buildings in VHR RSIs usually have clear shape patterns, it is meaningful to use the shape constraints to alleviate these problems. To this end, we propose the adversarial shape learning network (ASLNet) to explicitly model these shape constraints. In this section, we describe in detail the architecture, loss functions, and the CNN modules of our ASLNet.

### A. Network Architecture

Fig. 2 illustrates the architecture of the proposed ASLNet for building extraction, which consists of a segmentation network and a discriminator network. The segmentation network itself is capable of segmenting buildings, while the discriminator is employed to guide the training of the segmentation network. The segmentation network follows the classic encoder-decoder structure in literature papers [3], [40], [41]. The encoder network contains down-sampling operations to extract high-level semantic features from image local patches, whereas the decoder network recovers the spatial resolution of encoded features. The choice of the encoder network is not the focus of this work, thus we simply adopt the ResNet [42] as the feature encoder. It has been widely used for feature extraction in building segmentation [43], road segmentation

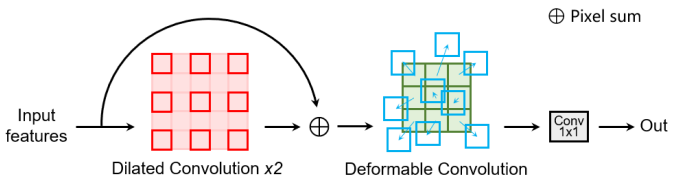



Fig. 3: The designed shape regularizer. Dilated convolutions and deformable convolutions are employed to enlarge the RFs and learn the shape features.

[44], and other semantic segmentation related tasks [45]. The selected ResNet version is ResNet34, which can be replaced by other versions based on the complexity of the dataset.

Apart from the output features from the late layers of the ResNet (with 1/8 of the original GSD), the early features (with 1/4 of the original GSD) are also employed in the decoder to learn finer spatial details. This is a commonly adopted design in segmentation networks [41], [40]. This ResNet with encoder-decoder structure is a modified version of FCN [2], denoted as ED-FCN. Compared with the plain FCN, the ED-FCN models the spatial features at a finer resolution, which is essential for the segmentation of VHR RSIs. It is therefore set as the baseline method of our segmentation network. Building on top of the ED-FCN, we further designed a shape regularizer at the end of the segmentation network in the proposed ASLNet to produce shape-refined outputs.

### B. Shape Regularizer

Although using a simple ResNet as the segmentation network is feasible for the adversarial shape learning, it is beneficial to model the shape features at finer spatial scales. Therefore, we design an explicit shape regularizer in the decoder of the segmentation network to enable a better adaptation to the shape constraints (see Fig. 3). The shape regularizer is placed at the spatial scale of 1/4 of the GSD, which operates on the fused multi-scale features in the ED-FCN. This spatial resolution for shape modeling is adopted following the practice in [41] and [40], which is a balance between accuracy and computational costs. At this spatial scale, a conventional $3 \times 3$ convolutional kernel has the RF of around $12 \times 12$ pixels, which is too small for modelling the local shape patterns. Therefore, we introduce the dilated convolution (DC) and deformable convolution (DFC) [46] layers to enlarge the RFs and to learn shape-sensitive transformations.

Both the DC and DFC are based on the idea of enlarging the coverage of convolutional kernels. Let us consider a convolutional operation for pixel $x(r, c)$ as:

$$U(r,c) = \sum_{i,j} x_{r+i,c+j} \cdot k_{i,j}, \tag{1}$$

where $k_{i,j}$ denotes the kernel weight. In a standard $3 \times 3$ convolution, $i, j \in \{-1, 0, 1\}$. However, in a $3 \times 3$ DC, $i, j \in \{-r, 0, r\}$ where $r$ is the dilation rate. In the designed SR, two $3 \times 3$ DCs are connected in a residual manner as in [11], composing a dilated residual unit. The residual branch allows the unit to gather information in different spatial ranges. In this way, the RF is enlarged to over $36 \times 36$ pixels.

A DFC is further employed to exploit the shape information, defined as:

$$U_{df}(r,c) = \sum_{i,j} x_{r+i+u(r,c),c+j+v(r,c)} \cdot k_{i,j}, \tag{2}$$

where $u(r, c)$ and $v(r, c)$ are position parameters learned by the additional convolutions, as follow:

$$u(r,c) = \sum_{i,j} x_{r+i,c+j} \cdot k^{'}_{i,j}, v(r,c) = \sum_{i,j} x_{r+i,c+j} \cdot k^{''}_{i,j}. \tag{3}$$

The DFC is placed at the end of the convolutional module (SR) as in [46]. This enables the SR to perceive and adapt to the local shape patterns. Finally, a $1 \times 1$ convolution is followed to project the learned features into a segmentation map.

### C. Shape Discriminator

A CNN model (even equipped with the SR) trained by the standard pixel-wise losses is not shape-aware, since each pixel is considered separately. To address this limit, we introduce a shape discriminator (SD) to drive the model to learn shape patterns. Although several literature works have introduced the adversarial learning for building extraction, most of them combine CNN outputs and input RSIs to train the discriminators [33], [30], [31], [32]. Under this condition, the discriminators are unlikely to learn the shape information, since they are affected by the redundant information in input RSIs. In the proposed ASLNet, the discriminator focuses only on the shape features, thus we exclude the use of input RSIs.

Training a shape discriminator with only binary inputs is challenging. Let $I$ denote an input image, $P$ be its corresponding prediction output and $L$ be the ground truth map. Since in $I$ there are usually mixed pixels (due to the sensor resolution) and discontinuities in objects representations (due to occlusions and low illumination conditions), it is common to have fuzzy areas in especially the building contours in the normalized prediction map $\sigma(P)$, where $\sigma$ is the Sigmoid function. However, in $L$ the human-annotated building contours have 'hard' edges, i.e. $L \in \{0, 1\}$. Mathematically, let $\sigma(P) \in [0, 1]$ be a smooth/fuzzy representation of the contours. This difference between $\sigma(P)$ and $L$ can be easily captured by the discriminator and causes failure to the shape modelling. In some literature works [29] a thresholding (or argmax) function $\mathcal{T}$ is employed to binarize $\sigma(P)$ as:

$$R = \mathcal{T}[\sigma(P)], \tag{4}$$

where R is the binary segmentation map. Although the obtained $R \in \{0, 1\}$, the $\mathcal{T}$ is non-differential in most cases, thus training the segmentation network with $R$ and $L$ will lead to zero-gradient problems.

In the designed shape discriminator we managed to eliminate this boundary difference and model only the shape information by adding a down-sampling operation $F_d$ in the discriminator $\mathcal{D}$. Fig. 4 illustrates the designed shape discriminator. After applying $F_d$, the building boundaries in $F_d(L)$ are 'softened' ($F_d(L) \in [0, 1]$) and the boundary difference

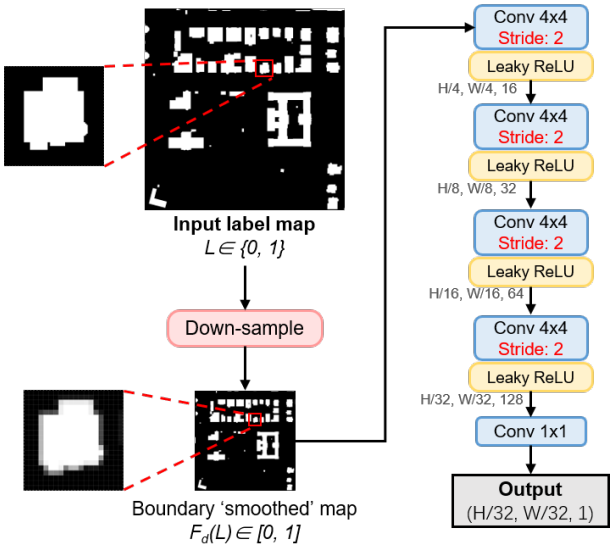



Fig. 4: The designed shape discriminator. The input maps are down-scaled to exclude the impact of 'hard' building boundaries in reference maps.

between $F_d(\sigma(P))$ and $F_d(L)$ is excluded. Specifically, four layers of strided convolution and activation functions are then employed to reduce the spatial size of feature maps and learn the local discriminative shape information. The output results are related to 1/32 of the original GSD.

The discriminator is trained with the Binary Cross Entropy (BCE) loss function. It is calculated as:

$$\begin{aligned}\mathcal{L}_{Dis} =& \mathbb{E}_{L\sim p_{\text{data}}(L)}[\log \mathcal{D}(L)] \\ &+\mathbb{E}_{P\sim p_{data}(P)}[\log(1-\mathcal{D}(\sigma(P)))] \\ =& -y\log(p)-(1-y)\log(1-p),\end{aligned} \tag{5}$$

where $\mathbb{E}$ is the expected value for different types of input samples, $y$ is the encoded signal that depending on the input map to the discriminator can be $L$ or $\sigma(P)$ ('1' and '0', respectively), and $p$ is the output of the discriminator. We employ the Mean Squared Error (MSE) loss function to calculate the $\mathcal{L}_{Shape}$ as:

$$\mathcal{L}_{Shape} = \{\mathcal{D}(L) - \mathcal{D}[\sigma(P)]\}^2, \tag{6}$$

where $\mathcal{D}$ is the shape discriminator. In this way, the $\mathcal{L}_{Shape}$ is related to the $L$, thus the segmentation network is constrained by the ground truth conditions.

### D. Optimization Objective of ASLNet

Let $\mathcal{L}_{Seg}$ be the loss function for the CNN-based segmentation of buildings. In conventional CNNs, $\mathcal{L}_{Seg}$ is only related to the pixel-wise accuracy, which does not consider the image context. In order to model the shape of objects with CNNs, it is essential to define a shape-based loss function $\mathcal{L}_{Shape}$. Previous works on shape analysis are often object-based [47], [48]. They include non-differential operations to calculate the shape measures, which are difficult to be incorporated into CNNs. Although there are also literature papers that use CNNs to regularize the shape of predictions [35], pre-training is often required and the regularization is limited to certain functions (e.g., inpainting of object contours). Since CNNs themselves can be trained to discriminate different shapes, we introduce the idea of adversarial learning to learn the $\mathcal{L}_{Shape}$ to guide the segmentation network.

$$\begin{aligned}\mathcal{L}_{Seg} =& \alpha\cdot\mathcal{L}_{Pix}+\beta\cdot\mathcal{L}_{Shape} \\ =& \alpha\cdot[L-\sigma(P)]^2+\beta\cdot\{\mathcal{D}(L)-\mathcal{D}[\sigma(P)]\}^2,\end{aligned} \tag{7}$$

where $\mathcal{L}_{Pix} = [L-\sigma(P)]^2$ is the supervised pixel-based reconstruction loss, $\alpha$ and $\beta$ are two weighting parameters. The first term in this formula drives the segmentation network to segment pixel-based $P$ in order to fit $L$, while the second term strengthens the local shape similarities between $P$ and $L$.

## IV. Design of Experiments

In this section, we describe the experimental dataset, the implementation details, and the considered evaluation metrics.

### A. Dataset Descriptions

We conduct building extraction experiments on two VHR RSI datasets, i.e., the Inria dataset [49] and the Massachusetts Building dataset [50]. These are two of the most widely studied building extraction datasets in the literature [6], [29], [14], [43].

*1) Inria Dataset [49]:* This is an aerial dataset with the GSD of 0.3 $m$ per pixel, covering 810 $km^2$. Each image has 5,000 $\times$ 5,000 pixels. There is a total of 360 images in this dataset, among which 180 are provided with the ground truth labels. These 180 images were collected in five different cities: Austin (U.S.), Chicago (U.S.), Kitsap (U.S.), Tyrol (Austria), and Vienna (Austria). Following the practice in [6], [14], we use the first 5 images in each city for testing and the rest 31 images for training.

*2) Massachusetts (MAS) Building Dataset [50]:* This is an aerial dataset collected on the Boston area. It has a GSD of 1.2 $m$ per pixel, covering around 340 $km^2$. The imaged regions include urban and suburban scenes where there are buildings with different sizes. This dataset consists of a training set with 137 images, a validation set with 4 images, and a test set with 10 images. Each image has 1,500 $\times$ 1,500 pixels.

### B. Implementation Details

The experiments were conducted on a workstation with 32 GB RAM and a NVIDIA Quadro P6000 GPU (23GB). Since it is impossible to train directly the large RSIs, they are randomly cropped into 512 $\times$ 512 patch images during the training process. The performed data preprocessing and augmentation operations include data normalization, random cropping, and image flipping. The training batch size is set to 8 and the number of training epochs is 50. The validation and test sets are evaluated on the original size RSIs to avoid the impact of cropping parameters. The hyper-parameters $\alpha, \beta$ in the Eq. (7) are set to 5.0, 1.0, respectively. The choice of hyper-parameters is discussed in Section V-A.

### C. Evaluation Metrics

*1) Pixel-based Evaluation Metrics:* We adopt several commonly used evaluation metrics in building extraction [6], [10] and other binary segmentation tasks [44] to assess the accuracy of the results. These metrics are based on statistical analysis of the classified pixels, including: overall accuracy ($OA$), Precision ($P$), Recall ($R$), F1 score, and mean Intersection over Union (IoU). The calculations are:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \tag{8}$$

$$F1 = 2 \times \frac{P \times R}{P+R}, \quad OA = \frac{TP+TN}{TP+FP+TN+FN}, \tag{9}$$

$$IoU = \frac{TP}{TP+FP+FN}, \tag{10}$$

where $TP$, $FP$, $TN$, and $FN$ represent true positive, false positive, true negative, and false negative, respectively.

*2) Object-based Evaluation Metrics:* Although the pixel-based evaluation metrics present the overall classification accuracy of the results, they fail to consider the thematic and geometrical properties of the segmented units [47]. To overcome this limitation, we designed three object-based evaluation metrics, including the matching rate ($MR$), the curvature error ($E_{curv}$), and the shape error ($E_{shape}$). These metrics are variants of the literature works [51], [48] to adapt to the assessment of building extraction results.

In order to compare the geometric quality of a segmented object $S_j$ on the prediction map $P$ and a reference object $O_i$ on the GT map $L$, it is essential to first discriminate if they are representing the same physical object. If $S_j$ and $O_i$ are overlapped, there are three possible overlapping relationships between them, as illustrated in Fig. 5. Therefore, for each $O_i$ $(i = 1, 2, 3, \cdots, n)$ and $S_j$ $(j = 1, 2, 3, \cdots, n')$, their matching relationship $M(O_i, S_j)$ is calculated based on the over-segmentation error ($E_{os}$) and under-segmentation error ($E_{us}$) [51]:

$$M(O_i, S_j) = \begin{cases} 0, & E_{os}(O_i, S_j) > T \,||\, E_{us}(O_i, S_j) > T \\ 1, & E_{os}(O_i, S_j) \leqslant T \,\&\, E_{us}(O_i, S_j) \leqslant T \end{cases} \tag{11}$$

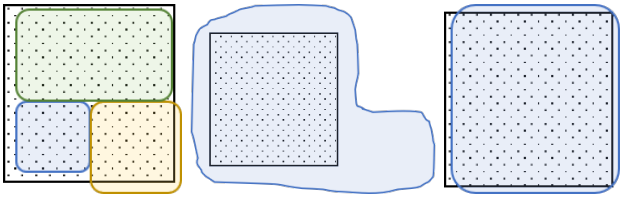
$$E_{os}(O_i, S_j) = 1 - \frac{|S_j \cap O_i|}{|O_i|}, \quad E_{us}(O_i, S_j) = 1 - \frac{|S_j \cap O_i|}{|S_j|}, \tag{12}$$

where $T$ is a threshold value (empirically set to 0.3). The matching rate ($MR$) of $P$ is the numeric ratio between the matched objects in $L$ and all the $O_i$ in $L$:

$$MR = \frac{\sum_{i,j} M(O_i, S_j)}{N_{O_i}}. \tag{13}$$

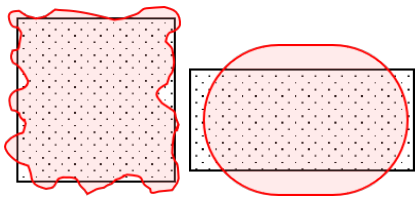
After finding the matched item $M_i$ in $P$ for $O_i$, two geometric measurements are further calculated to measure the differences between $M_i$ and $O_i$. First, $E_{curv}$ is introduced to measure the differences in object boundaries. It is calculated as:

$$E_{curv}(O_i, M_i) = ||f_c(M_i) - f_c(O_i)||, \tag{14}$$



(a) (b) (c)

Fig. 5: Illustration of three overlapping relationships between a segmented object $S_j$ (colored region) and a reference object $O_i$ (dotted region). (a) Over-segmentation, (b) Under-segmentation, and (c) Matching.



(a) (b)

Fig. 6: Examples of the reference object $O_i$ (dotted region) and its matched segmented object $M_i$ (colored region) that have: (a) high curvature error ($E_{curv}$), and (b) high shape error ($E_{shape}$).

where $f_c$ denotes the contour curvature function [52]. Since $O_i$ is human-annotated, $f_c(O_i)$ is usually small. A large $E_{curv}(O_i, M_i)$ indicates that the boundary of $f_c(M_i)$ is uneven. The second measurement $E_{shape}$ is introduced to assess the difference in shape, calculated as:

$$E_{shape}(O_i, M_i) = ||f_s(M_i) - f_s(O_i)||, \quad f_s(M_i) = \frac{4\pi|M_i|}{p_{M_i}^2}, \tag{15}$$

where $p_{M_i}$ is the perimeter of $M_i$. The value of $f_s(M_i)$ is 1 for a circle and $pi/4$ for a square [48], [52]. Two examples of the curvature and shape errors are illustrated in Fig. 6.

## V. Experimental Results

This section presents the experimental results obtained on the two VHR building datasets. First, we present the ablation study to quantitatively evaluate the improvements brought by the proposed method. Then the effects of the shape regularizer (SR) and the shape discriminator (SD) are analyzed in greater detail on some significant sample areas. Finally, the proposed ASLNet is compared with several state-of-the-art CNN models for building extraction.

### A. Ablation Study

**Influence of Hyper-Parameters.** The hyper-parameters $\alpha$ and $\beta$ in Eq. (7) balance $\mathcal{L}_{pix}$ and $\mathcal{L}_{shape}$. To find which set of hyper-parameters leads to the best performance, we conduct an experiment on the Inria dataset. We set the value of one of the parameters to 1 and change the other one. The mIoU obtained with different hyper-parameter values are reported in Table I. We find that setting $\mathcal{L}_{pix}$ as the primary loss (i.e.,

| Hyper-parameter | 1 | 3 | 5 | 10 |
|---|---|---|---|---|
| $\alpha(\beta = 1)$ | 77.56 | 78.58 | **79.30** | 78.82 |
| $\beta(\alpha = 1)$ | 77.56 | 76.00 | 75.21 | 65.97 |

TABLE I: The mIoU under different hyper-parameters tested on the Inria dataset.

| Adversarial Loss | OA(%) | F1(%) | mIoU(%) |
|---|---|---|---|
| BCE | 96.67 | 86.23 | 76.26 |
| BCE + FM [53] | 96.20 | 84.67 | 73.81 |
| Perceptual [54] | 96.09 | 84.38 | 73.35 |
| Multi-scale L1 [32] | 96.45 | 85.67 | 75.39 |
| MSE (adopted) | **97.15** | **88.27** | **79.30** |

TABLE II: The accuracy obtained by training with different adversarial losses on the Inria dataset.

$\alpha > \beta$) leads to higher accuracy. The ASLNet obtains the best accuracy when $\alpha = 5, \beta = 1$. Therefore, these hyper-parameters are fixed in adversarial training of the ASLNet in all the experiments.

**Choice of Adversarial Losses.** There are a variety of loss functions available for the adversarial training. We test some of the commonly used losses for our task, including: 1) *BCE loss*. It is calculated between the outputs of the discriminator and the domain labels (i.e., whether inputs to the discriminator are predictions or GT maps); 2) *Feature Matching (FM) loss* [53]. It is an auxiliary loss commonly used to stabilize the training of GANs. It matches the moments of the activation on an intermediate layer of the discriminator; 3) *Perceptual loss* [54]. It calculates the distance between features extracted from generated and GT images using a pretrained network; 4) *Multi-scale L1 loss* [32]. It calculates the L1 distance of features in the discriminator extracted from the prediction and GT maps; and 5) *MSE loss*. It is calculated as in Eq. (6).

The obtained accuracy is reported in Table II. The BCE loss (either w/ or w/o auxiliary loss) causes training instability and leads to unsatisfactory results, as it encourages the segmentation network to generate fake predictions unrelated to the GT situations. The perceptual loss drives the segmentation network to pay more attention to the boundary of objects (instead of the shape), since the pretrained network is not sensitive to shape features. On the contrary, the multi-scale L1 loss aligns only the features without considering the segmentation maps, thus the trained network fails to optimize the building boundaries. The MSE loss successfully drives the segmentation network to learn shape patterns, leading to the highest accuracy. Therefore, it is adopted as the $\mathcal{L}_{Shape}$ to train the segmentation network.

**Quantitative Results.** We conduct extensive ablation studies to assess the effectiveness of the proposed ASLNet. To compare the results before and after the use of SR and SD, the original FCN [2] and the baseline method ED-FCN are also included in the comparison. The quantitative results are reported in Table III. The baseline ED-FCN outperforms the FCN in terms of mean IoU by 0.21% and 4.87%, respectively in the Inria and the MAS dataset, which is attributed to the concatenation of low-level features in its decoder. Since the MAS dataset has lower spatial resolution, the improvements of the ED-FCN is more noticeable. The ASLNet w/ the SR but w/o the SD has slight accuracy improvements over the ED-FCN. Meanwhile, after introducing the adversarial shape learning, the ASLNet (w/o the SR) has the mean IoU improvements of 1.56% and 2.63% on the two datasets. The complete ASLNet with both the SR and the SD provides improvements of 2.73% and 3.26% in mean IoU compared to the baseline ED-FCN. Fig. 9 shows a comparison of the OA values of the segmented probability maps versus different binarization ($\mathcal{T}$ in Eq. (4)) thresholds. Since the ASLNet directly segments near-binary regularized results, its OA curves are close to horizontal, and are sharply above the baseline methods.

The improvements are even more significant in terms of object-based metrics. The baseline FCN encountered severe over-segmentation problems, which lead to low $MR$ values. The ED-FCN and the ASLNet (w/o the SD) slightly improve the three object-based metrics. The ASLNet (w/o the SR) has improvements of around 3% in both $E_{curv}$ and $E_{shape}$ in the two datasets. The complete ASLNet further improves the $MR$ values of around 4% on the two datasets.

**Qualitative Results.** Fig. 7 shows the results of the ablation study on several sample areas. The segmentation results of the ED-FCN are generally round-edged. However, after adding the SD, the building edges became sharper and the object shapes became more rectangular. Moreover, the object shapes are modelled in a wider image range, thus the edges are more straight and some missing parts are inpainted. More specifically, Fig. 7(a) and Fig. 7(e) show two cases of occlusions caused by trees and shadows, respectively. Fig. 7(c) shows a case of under-segmentation. In these cases the ASLNet has successfully recovered the complete buildings. Fig. 7(b), (d), and (f) show several examples of the improvements in shapes. It is worth noting that the ASLNet managed to improve the segmented shape of compact small objects (e.g., houses), irregular large object (e.g., factories), and long bar-like objects (e.g., residential buildings). However, a side-effect of the ASLNet is that it fails to segment some round objects (e.g., oil tanks) that are unseen in the training set. The learned shape bias drives the ASLNet to optimize the rectangular contour of buildings. Some of examples of these cases are shown in Fig. 8. Considering the objective of the proposed method, this drawback has minor impacts. Note that the proposed shape-driven training could also be adapted to other general shapes to suit different applications.

As a conclusion of the ablation study, the modeling of shape features in the ASLNet leads to three significant benefits: 1) inpainting of the missing parts of buildings; 2) providing a joint segmentation and regularization of the building contours; 3) mitigating the under-segmentation and over-segmentation problems. These advantages are verified by both the accuracy metrics and visual observation.

**ASLNet with SOTA Backbones.** For assessing the performance of the proposed techniques, the ASLNet is designed on top of a simple ED-FCN. However, replacing the ED-FCN with more advanced segmentation networks may potentially improve its accuracy. To test this, we integrate the SR and SD modules into two well-known and widely used segmentation backbones, i.e., the DeepLabv3+ [41] and the HRNet [55].

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3134455



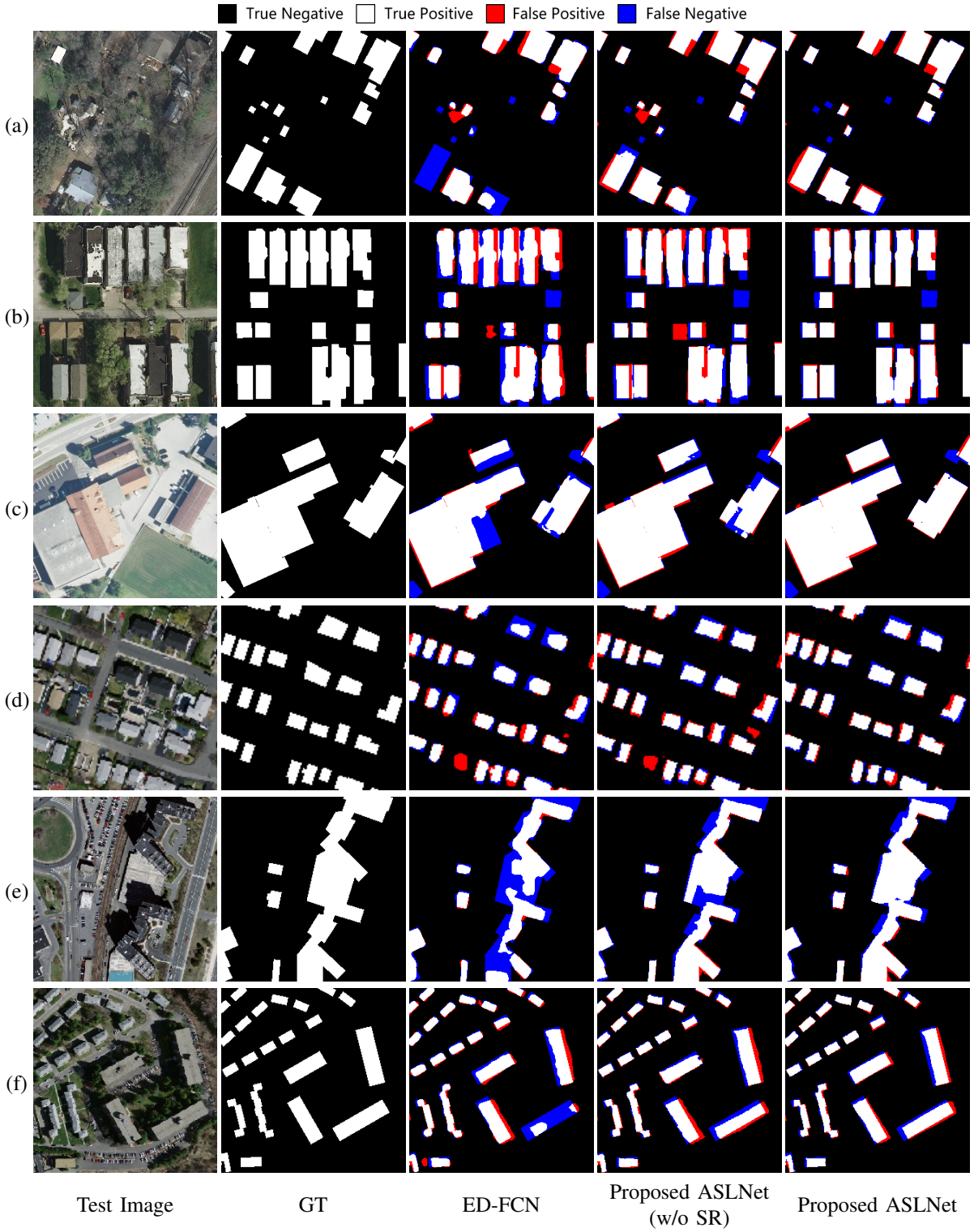



Fig. 7: Examples of segmentation results obtained by the different methods (ablation study). (a)-(c) Results selected from the Inria dataset, (d)-(f) Results selected from the Massachusetts dataset.

The SR module is placed at the end of each segmentation backbone, while the SD module is used in the same way as in the original ASLNet. The resulting variants of the ASLNet are referred to as the ASLNet-DL and the ASLNet-HR, respectively.

The quantitative results are reported in Table IV. Both ASLNet-DL and ASLNet-HR obtain sharp accuracy improvements over their baselines (DeepLabv3+ and HRNet), proving that the proposed shape training method is effective on different segmentation backbones. Compared to the ASLNet, the ASLNet-HR obtains slight accuracy improvements on the two

TABLE III: Results of the ablation study on the two considered data sets.

| Dataset | Method | Components | | Pixel-based Metrics | | | | | Object-based Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SR | SD | OA(%) | P(%) | R(%) | F1(%) | mIoU(%) | $MR$(%) | $E_{curv}$ | $E_{shape}$ |
| Inria | FCN [2] | | | 96.72 | 89.41 | 83.78 | 86.33 | 76.36 | 55.37 | 7.66 | 6.63 |
| | ED-FCN | | | 96.69 | 87.87 | 85.29 | 86.46 | 76.57 | 60.38 | 7.26 | 6.29 |
| | Proposed ASLNet (w/o SD) | √ | | 96.71 | 87.18 | 86.77 | 86.82 | 77.06 | 57.36 | 7.42 | 6.21 |
| | Proposed ASLNet (w/o SR) | | √ | 96.94 | 88.98 | 86.32 | 87.50 | 78.13 | 60.36 | 3.86 | 4.36 |
| | Proposed ASLNet | √ | √ | **97.15** | **90.00** | **86.85** | **88.27** | **79.30** | **64.46** | **3.53** | **3.66** |
| MAS | FCN [2] | | | 92.39 | 78.46 | 78.73 | 78.56 | 64.82 | 26.87 | 11.56 | 7.79 |
| | ED-FCN | | | 93.81 | 84.83 | 79.57 | 82.09 | 69.69 | 53.62 | 8.78 | 7.45 |
| | Proposed ASLNet (w/o SD) | √ | | 93.95 | 85.47 | 79.45 | 82.31 | 70.03 | 55.04 | 8.69 | 7.11 |
| | Proposed ASLNet (w/o SR) | | √ | 94.38 | 85.70 | 81.17 | 83.91 | 72.32 | 62.39 | 7.36 | 4.30 |
| | Proposed ASLNet | √ | √ | **94.51** | **85.92** | **82.83** | **84.32** | **72.95** | **67.28** | **7.19** | **4.01** |

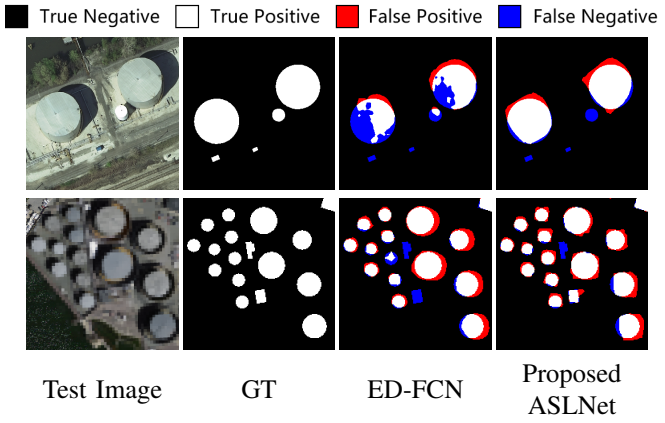
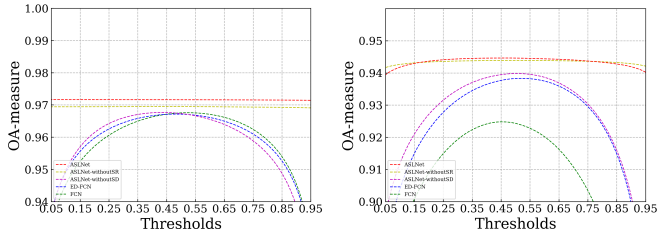



Fig. 8: Examples of the failure cases. The ASLNet segments rectangular items for even the round objects, given its building-shape driven training.




(a) (b)

Fig. 9: Accuracy curves versus different binarization threshold of (a) Inria dataset, and (b) Massachusetts dataset.

considered datasets, whereas accuracy of the ASLNet-DL is decreased on the MAS dataset. This suggests that the atrous convolutions operated on high-level features is not effective on the MAS dataset (which has a relatively lower GSD).

### *B. Comparative Experiments*

**Quantitative Results.** We further compare the proposed ASLNet with several literature works to assess its effectiveness. Three classic models for the semantic segmentation are compared, including the UNet [3], the baseline method FCN [2] and the Deeplabv3+ [41]. The cwGAN-gp [33] that uses GAN for building extraction is also compared. Moreover, we compare the proposed method with several state-of-the-art methods for building extraction, including the ResUNet [4], the MAPNet [5], the GMEDN [14] and the FC-DenseNet+FRCRF [16] (which includes a CRF-based post-processing step). The quantitative results on the Inria dataset and the MAS dataset are reported in Table V and Table VI, respectively.

TABLE IV: Results obtained using SOTA segmentation backbones.

| Dataset | Method | Backbone | OA(%) | F1(%) | mIoU(%) |
|---|---|---|---|---|---|
| Inria | DeepLabv3+ [41] | DeepLabv3+ | 96.85 | 86.97 | 77.30 |
| | ASLNet-DL (Proposed) | | **97.18** | **88.26** | **79.31** |
| | HNRet [55] | HNRet | 96.90 | 87.18 | 77.68 |
| | ASLNet-HR (Proposed) | | **97.20** | **88.40** | **79.54** |
| MAS | DeepLabv3+ [41] | DeepLabv3+ | 93.27 | 80.53 | 67.52 |
| | ASLNet-DL (Proposed) | | **94.41** | **83.88** | **72.31** |
| | HNRet [55] | HNRet | 94.34 | 83.33 | 71.55 |
| | ASLNet-HR (Proposed) | | **94.61** | **85.00** | **73.99** |

Let us first analyze the pixel-based metrics. The ResUNet, which is a variant of UNet for the building extraction, outperforms the classic semantic segmentation models (UNet, FCN and Deeplabv3+) by a large margin on the MAS dataset. The accuracy of cw-GAN-gp is higher than that of the FCN on the MAS dataset but it is lower on the Inria dataset. on the The MAPNet obtains competitive results on the Inria dataset, whereas its performance is inferior to the ResUNet and the Deeplabv3+ on the MAS dataset. On the contrary, the GMEDN obtains better accuracy on the MAS dataset. The FCN-DenseNet+FRCRF achieves the second best accuracy on the MAS dataset. The proposed ASLNet outperforms all the compared methods in almost all the metrics (except for the precision and recall on the MAS dataset), although its baseline method (the ED-FCN) is inferior to most of them. The advantages of the ASLNet are particularly noticeable on the Inria dataset, where the ASLNet improves the mean IoU of 1.51% with respect to the second best method. The reason for which the ASLNet has higher improvements on the Inria dataset can be attributed to the higher GSD of this dataset, where the building shape information is more discriminative.

In terms of object-based metrics, there are remarkable differences in the $MR$ values. The cw-GAN-gp and the ResUNet obtained the third best $MR$ values among the literature methods on the Inria dataset and the MAS dataset, respectively. The FCN-DenseNet+FRCRF obtained the second-best accuracy in all the object-based metrics due to its boundary-refinement CRF operations. All the other compared literature

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3134455



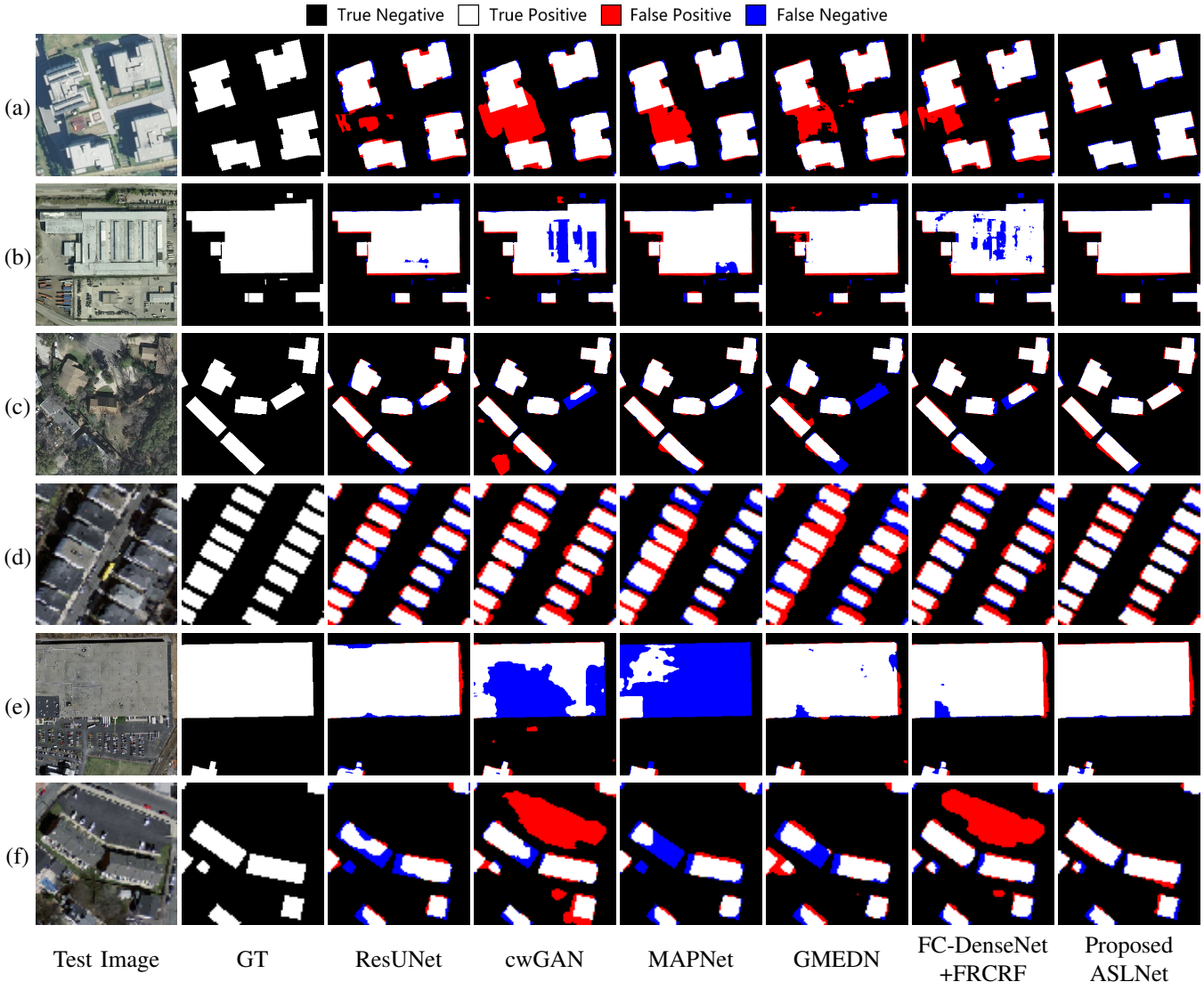



Fig. 10: Examples of segmentation results obtained by the different methods (comparative experiments). (a)-(c) Results selected from the Inria dataset, (d)-(f) Results selected from the Massachusetts dataset.

methods obtained very high $E_{curv}$ and $E_{shape}$ values. This indicates that they all suffer from irregular shapes and uneven boundaries problems. On the contrary, the proposed ASLNet shows significant advantages in terms of all these three metrics. Due to its learned shape constraints that regularize the segmented items and sharpen the building boundaries, the ASLNet exhibits great advantages in $E_{shape}$ and $E_{curv}$ in both datasets.

**Qualitative Results.** Fig. 10 shows comparisons of the segmentation results obtained by the compared methods. One can observe that the proposed ASLNet exhibits several advantages in different scenes. It is capable of accurately segmenting the individual buildings in Fig. 10(a), the occluded houses in Fig. 10(c) and the large-size factories/supermarkets in Fig. 10(b) and Fig. 10(e). When it deals with dense residential buildings as shown in Fig. 10(d), the over-segmentation and under-segmentation errors are reduced. It also excludes some uncertain areas by considering the shape patterns (e.g., the colored opening space in Fig. 10(a) and the parking lot in Fig. 10(f)).

## VI. Conclusions

Recent works on CNN-based building extraction exhibit severe limitations resulting in two main issues: 1) incomplete segmentation of objects due to occlusions and intra-class diversity; 2) geometric regularization of the building extraction results. To address these issues, we introduce the adversarial training strategy to learn the shape of buildings and propose an ASLNet. Specifically, we designed a SR with shape-sensitive convolutional layers (DCs and DFCs) to regularize the feature maps, as well as a SD to learn the shape constraints to guide the segmentation network. The SR and SD allow an accurate modelling of the shape information contained in the considered images. To the best of our knowledge, this is the first work that learns adversarial shape constraints for the segmentation of RSIs. To quantitatively evaluate the thematic properties of

TABLE V: Results of the comparative experiments on the Inria dataset.

| Method | Pixel-based Metrics | | | | | Object-based Metrics | | |
|---|---|---|---|---|---|---|---|---|
| | OA(%) | P(%) | R(%) | F1(%) | mIoU(%) | $MR$(%) | $E_{curv}$ | $E_{shape}$ |
| UNet [3] | 95.52 | 81.76 | 82.76 | 82.03 | 70.03 | 43.87 | 10.89 | 7.84 |
| FCN [2] | 96.72 | 89.41 | 83.78 | 86.33 | 76.36 | 55.37 | 7.66 | 6.63 |
| Deeplabv3+ [41] | 96.85 | 89.17 | 85.09 | 86.97 | 77.30 | 58.63 | 7.12 | 6.29 |
| ResUNet [4] | 96.50 | 88.33 | 83.60 | 85.68 | 75.41 | 55.72 | 7.47 | 6.50 |
| cwGAN-gp [33] | 96.54 | 86.43 | 85.61 | 85.94 | 75.76 | 61.51 | 7.10 | 5.47 |
| MAPNet [5] | 96.96 | 88.58 | 86.04 | 87.24 | 77.79 | 59.75 | 6.26 | 6.16 |
| GMEDN [14] | 96.23 | 87.03 | 81.37 | 83.88 | 72.95 | 52.65 | 8.43 | 5.54 |
| FC-DenseNet+FRCRF [16] | 96.74 | 89.55 | 83.68 | 86.36 | 76.34 | 63.43 | 4.31 | 4.13 |
| ASLNet (proposed) | **97.15** | **90.00** | **86.85** | **88.27** | **79.30** | **64.46** | **3.53** | **3.66** |

TABLE VI: Results of the comparative experiments on the Massachusetts dataset.

| Method | Pixel-based Metrics | | | | | Object-based Metrics | | |
|---|---|---|---|---|---|---|---|---|
| | OA(%) | P(%) | R(%) | F1(%) | mIoU(%) | $MR$(%) | $E_{curv}$ | $E_{shape}$ |
| UNet [3] | 92.18 | 84.71 | 70.29 | 76.75 | 62.34 | 40.02 | 10.23 | 7.10 |
| FCN [2] | 92.39 | 78.46 | 78.73 | 78.56 | 64.82 | 26.87 | 11.56 | 7.79 |
| Deeplabv3+ [41] | 93.27 | 82.28 | 78.95 | 80.53 | 67.52 | 47.15 | 9.82 | 7.67 |
| ResUNet [4] | 94.32 | 86.16 | 81.25 | 83.59 | 71.87 | 60.22 | 7.91 | 7.16 |
| cw-GAN-gp [33] | 93.00 | 81.03 | 79.64 | 80.29 | 67.15 | 51.94 | 9.37 | 6.74 |
| MAPNet [5] | 93.47 | **87.88** | 72.77 | 79.50 | 66.20 | 53.70 | 8.05 | 7.63 |
| GMEDN [14] | 93.29 | 84.09 | 77.49 | 80.63 | 67.61 | 51.20 | 9.20 | 7.26 |
| FC-DenseNet+FRCRF [16] | 94.48 | 85.28 | **83.16** | 84.18 | 72.77 | 67.21 | 7.92 | 6.66 |
| ASLNet (proposed) | **94.51** | 85.92 | 82.83 | **84.32** | **72.95** | **67.28** | **7.19** | **4.01** |

the building extraction results, we also designed three object-based metrics: the matching rate, the curvature error and the shape error.

Experimental results on two VHR building datasets show that the proposed ASLNet has obtained significant improvements over the conventional CNN models in both pixel-based metrics and object-based metrics. These improvements can be attributed to two factors. First, learning the shape priors is beneficial to inpaint the missing building parts. Second, the shape constraints force the ASLNet to produce shape-regularized results, thus the segmented objects have rectangular shape and smooth boundaries. Additionally, we observed that the ASLNet greatly reduces the over-segmentation and under-segmentation errors (proved by the higher $MR$ values). One of the limitation of the ASLNet is that it reduces its accuracy on the segmentation of objects with shape that are not rectangular (e.g., round buildings), which is due to its learned shape constraints.

The adversarial shape learning is potentially beneficial for other segmentation-related tasks with the RSIs, where the ground objects exhibit certain geometric patterns. In future studies, we will investigate to use the adversarial shape learning to model other types of object shapes in different tasks (e.g., road extraction, change detection and land-cover mapping in RSIs).

## REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, vol. 39, no. 12, pp. 2481–2495, 2017.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[4] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, no. 1, p. 144, 2018.

[5] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[6] Y. Xie, J. Zhu, Y. Cao, D. Feng, M. Hu, W. Li, Y. Zhang, and L. Fu, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1842–1855, 2020.

[7] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.

[8] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2793–2798, 2017.

[9] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the united states," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2600–2614, 2018.

[10] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 184–197, 2020.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[12] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set,"

*IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[13] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[14] J. Ma, L. Wu, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Building extraction of aerial images by a global and multi-scale encoder-decoder network," *Remote Sensing*, vol. 12, no. 15, p. 2350, 2020.

[15] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask r-cnn with building boundary regularization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 247–251.

[16] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf)," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7502–7519, 2020.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[18] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[19] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *ECCV*, 2020.

[20] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *ICCV*, 2019.

[21] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Gesturegan for hand gesture-to-gesture translation in the wild," in *ACM MM*, 2018.

[22] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *CVPR*, 2018.

[23] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, 2019.

[24] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *CVPR*, 2017.

[25] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in *CVPR*, 2017.

[26] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *CVPR*, 2018.

[27] L. Tran, K. Sohn, X. Yu, X. Liu, and M. Chandraker, "Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild," in *CVPR*, 2019.

[28] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial cross-domain action recognition with co-attention," in *AAAI*, 2020.

[29] X. Li, X. Yao, and Y. Fang, "Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3680–3687, 2018.

[30] B. Bischke, P. Helber, F. Koenig, D. Borth, and A. Dengel, "Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2018, pp. 1–6.

[31] A. Abdollahi, B. Pradhan, S. Gite, and A. Alamri, "Building footprint extraction from high resolution aerial images using generative adversarial network (gan) architecture," *IEEE Access*, vol. 8, pp. 209 517–209 527, 2020.

[32] X. Pan, F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren, "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sensing*, vol. 11, no. 8, p. 917, 2019.

[33] Y. Shi, Q. Li, and X. X. Zhu, "Building footprint generation using improved generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 4, pp. 603–607, 2018.

[34] H. A. Atabay, "Binary shape classification using convolutional neural networks," *IIOAB J*, vol. 7, no. 5, pp. 332–336, 2016.

[35] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 203–211.

[36] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5229–5238.

[37] W. Kuo, A. Angelova, J. Malik, and T.-Y. Lin, "Shapemask: Learning to segment novel objects by refining shape priors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9207–9216.

[38] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic correlation promoted shape-variant context for segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8885–8894.

[39] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, "Polytransform: Deep polygon transformer for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9131–9140.

[40] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[41] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[43] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sensing*, vol. 11, no. 7, p. 830, 2019.

[44] L. Ding and L. Bruzzone, "Diresnet: Direction-aware residual network for road extraction in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[45] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[46] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

[47] S. Ye, R. G. Pontius Jr, and R. Rakshit, "A review of accuracy assessment for object-based image analysis: From per-pixel to per-polygon approaches," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 141, pp. 137–147, 2018.

[48] I. Lizarazo, "Accuracy assessment of object-based image classification: another step," *International Journal of Remote Sensing*, vol. 35, no. 16, pp. 6135–6156, 2014.

[49] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[50] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

[51] C. Persello and L. Bruzzone, "A novel protocol for accuracy assessment in classification of very high resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1232–1244, 2009.

[52] R. C. Gonzalez and R. E. Woods, "Digital image processing. upper saddle river," *J.: Prentice Hall*, 2002.

[53] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

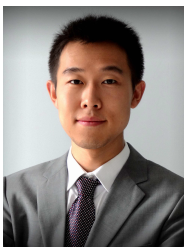
[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[55] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

**Lei Ding** received the MS's degree in Photogrammetry and Remote Sensing from the Information Engineering University (Zhengzhou, China), and the PhD (cum laude) in Communication and Information Technologies from the University of Trento (Trento, Italy). His research interests are related to semantic segmentation, change detection and domain adaptation with Deep Learning techniques. He is a referee for many international journals, including IEEE TIP, TNNLS, TGRS, GRSL and JSTAR.

**Hao Tang** is currently a Postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He received the master's degree from the School of Electronics and Computer Engineering, Peking University, China and the Ph.D. degree from Multimedia and Human Understanding Group, University of Trento, Italy. He was a visiting scholar in the Department of Engineering Science at the University of Oxford. His research interests are deep learning, machine learning, and their applications to computer vision.

**Yahui Liu** is a Ph.D. student in the Department of Information Engineering and Computer Science at the University of Trento, Italy. Before that, he received his B.S. degree and M.S. degree from Wuhan University, China, in 2015 and 2018, respectively. His major research interests are machine learning and computer vision, including unsupervised learning and image domain translation.



**Yilei Shi** (M'18) received his Diploma (Dipl.-Ing.) degree in Mechanical Engineering, his Doctorate (Dr.-Ing.) degree in Engineering from Technical University of Munich (TUM), Germany. In April and May 2019, he was a guest scientist with the department of applied mathematics and theoretical physics, University of Cambridge, United Kingdom. He is currently a senior scientist with the Chair of Remote Sensing Technology, Technical University of Munich.

His research interests include computational intelligence, fast solver and parallel computing for large-scale problems, advanced methods on SAR and InSAR processing, machine learning and deep learning for variety data sources, such as SAR, optical images, medical images and so on; PDE related numerical modeling and computing.

**Xiao Xiang Zhu** (S'10–M'12–SM'14–F'21) received the Master (M.Sc.) degree, her doctor of engineering (Dr.-Ing.) degree and her "Habilitation" in the field of signal processing from Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She is currently the Professor for Data Science in Earth Observation (former: Signal Processing in Earth Observation) at Technical University of Munich (TUM) and the Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019, Zhu is a co-coordinator of the Munich Data Science Research School (www.muds.de). Since 2019 She also heads the Helmholtz Artificial Intelligence – Research Field "Aeronautics, Space and Transport". Since May 2020, she is the director of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond", Munich, Germany. Since October 2020, she also serves as a co-director of the Munich Data Science Institute (MDSI), TUM. Prof. Zhu was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan and University of California, Los Angeles, United States in 2009, 2014, 2015 and 2016, respectively. She is currently a visiting AI professor at ESA's Phi-lab. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing and serves as the area editor responsible for special issues of IEEE Signal Processing Magazine. She is a Fellow of IEEE.

**Lorenzo Bruzzone** (S'95-M'98-SM'03-F'10) received the Laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications. Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is the Principal Investigator of the *Radar for icy Moon exploration* (RIME) instrument in the framework of the *JUpiter ICy moons Explorer* (JUICE) mission of the European Space Agency. He is the author (or coauthor) of 215 scientific publications in referred international journals (154 in IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. He was invited as keynote speaker in more than 30 international conferences and workshops. Since 2009 he is a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS).

Dr. Bruzzone was a Guest Co-Editor of many Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003 he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the founder of the IEEE Geoscience and Remote Sensing Magazine for which he has been Editor-in-Chief between 2013-2017. Currently he is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing. He has been Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society between 2012-2016. His papers are highly cited, as proven form the total number of citations (more than 27000) and the value of the h-index (78) (source: Google Scholar).